GENOME ANALYSIS

Statement as to Federally Sponsored Research

Funding for the work described herein was provided by the federal government, which may have certain rights in the invention.

BACKGROUND

1. Technical Field

The invention relates to methods and materials involved in the analysis of an organism's genome. Specifically, the invention relates to methods and materials for identifying genomic markers, mapping genomic markers, and identifying genomic sequences that contribute to specific phenotypic traits.

2. Background Information

In general, the genome of an organism controls that organism's phenotype. Thus, understanding the organization and function of an organism's genome can allow scientists to manipulate particular traits. For example, a greater understanding of the organization and function of the maize genome is essential to enhance the efficiency and effectiveness of breeding programs designed to meet the growing needs for maize as food, feed, and industrial feedstocks.

In an attempt to understand genomic organization, genome sequencing projects have been initiated. In fact, the complete sequences of over a dozen genomes have been obtained in the last few years. While such projects provide much useful organizational information, limited functional information is obtained. For example, one of the most surprising results from these analyses has been the large percentage (typically 30-40%) of novel genes discovered for which no molecular function can be assigned via sequence comparisons. Thus, aside from being almost prohibitively expensive, genome sequencing projects by themselves fail to provide optimal functional information to aid genetic modification efforts.

Briefly, alleles of single genes are responsible for the discrete phenotypic classes that are observed in families segregating for Mendelian mutants. Many of the phenotypes of economic significance in humans, livestock, and plants, however, are "quantitative traits." For example, traits such as susceptibility to heart disease in

humans, litter size in pigs, and yield in maize are controlled by many genetic loci working in concert. As such, these traits exhibit continuous variation and are often highly susceptible to pronounced environmental interactions. Consequently, it has been difficult to obtain an understanding of the molecular basis of important traits of this type. Nevertheless, plant breeders have been successful at developing empirically validated selection methods. Indeed, the average annual rate of genetic gain for maize yields during the past 60 years has been 1.5 percent. There is still, however, only a very limited understanding of the molecular mechanisms responsible for high, stable yields. Hence, the ability of breeders to identify superior germplasm prior to field testing and to improve selection practices remains limited. This is of great concern given that it appears that the rate of genetic gain in maize breeding programs has been leveling off during the last two decades and because plant breeders now face new environmental and scientific challenges.

Thus, two of the most significant challenges that biologists face in using genomic data to manipulate particular traits are: 1) assigning functions to novel genes; and 2) understanding the molecular basis of, for example, quantitative genetics and heterosis.

SUMMARY

The invention involves methods and materials related to the analysis of an organism's genome. Specifically, the invention provides methods and materials for identifying genomic markers, mapping genomic markers, and identifying genomic sequences that contribute to specific traits. For example, the invention provides methods and materials that can be used to assign functions to genes by genetically mapping a large collection of nucleic acid fragments (e.g., cDNAs). These methods and materials also can be used to facilitate the genetic mapping of genes responsible for the large collection of Mendelian mutants from any species (e.g., maize). By so doing, it will make it possible to associate mutant phenotypes with small numbers of sequence-defined genes (i.e., candidate gene cloning). In addition, the invention provides methods and materials that can be used to dissect molecularly a genome (e.g., a maize genome) and identify chromosomal regions and specific groups of genes that influence quantitative traits and heterosis.

Further, the invention provides methods and materials that can be used 1) to develop a dense genetic map populated with a novel class of markers (insertion/deletion polymorphisims; IDPs) that can be used in allele-specific, high-throughput analyses; 2) to map genetically a large number of non-redundant, sequence-defined nucleic acid fragments (e.g., cDNAs); 3) to map genetically, with high resolution, genes responsible for specific mutant phenotypes, thereby associating mutant phenotypes with small numbers of sequence-defined genes; 4) to identify via high-throughput, allele-specific, IDP markers, chromosomal intervals that have undergone alterations in allele frequencies in economically significant populations (e.g., maize populations that have been selected over the last 50 years for increased levels of grain yield and heterosis); and 5) to identify, via a microarray technology, genes whose patterns of expression are controlled by selected chromosomal intervals or altered in F_1 hybrids relative to their parental inbreds.

In general, the invention features an array containing a nucleic acid component consisting essentially of non-redundant nucleic acid molecules. The array may contain at least about 50 percent, at least about 75 percent, at least about 90 percent, or at least about 95 percent, of the non-redundant nucleic acid molecules corresponding to an untranslated sequence in an organism. In addition, the array may contain at least about 50 percent of the non-redundant nucleic acid molecules corresponding to a 3' untranslated sequence in an organism, or at least about 50 percent of the non-redundant nucleic acid molecules corresponding to a 5' untranslated sequence in an organism, or at least about 50 percent of the non-redundant nucleic acid molecules corresponding to an intronic sequence in an organism. The array may contain more than about 500, or more than about 1000, of the non-redundant nucleic acid molecules. Further, the sequence of each non-redundant nucleic acid molecule may be known. A representative organism is a plant, and a representative plant is a corn plant.

In addition, an array of the invention containing the non-redundant nucleic acid molecules may have nucleic acid sequences corresponding to different sequences transcribed in a cell. The nucleic acid component may contain at least two groups of non-redundant nucleic acid molecules, wherein each non-redundant nucleic acid molecule within each group has a nucleic acid sequence corresponding to a different

sequence transcribed in a cell from a source, with the source being different for each group. The array may contain at least ten groups. In addition, each non-redundant nucleic acid molecule may have a marker such that the source is identifiable. Representative markers include nucleic acid markers. The source may be an organ tissue at a stage of development. Representative organ tissues include roots, shoots, stems, leaves, flowers, seeds, or meristems, and representative developmental stages are germinating seedlings, full-grown plants, and immature/developing seeds.

In general, another feature of the invention is an IDP primer pair collection having at least about 100 different IDP primer pairs. The first primer of each of the IDP primer pair typically corresponds to a different first sequence within the genome of at least one member of a species, each different first sequence lacking an IDP for the species, wherein the second primer of each of the IDP primer pairs corresponds to a different second sequence within the genome of at least one member of the species, each different second sequence containing an IDP for the species. The collection may include at least about 250, at least about 500, or at least about 1000 different IDP primer pairs. In addition, the sequence of each primer may be known. It is a feature of the collection of IDP primer pairs that every fifty cM region, every twenty-five cM region, every ten cM region, every five cM region, or every two cM region of the genome contains at least one of the different first sequences.

It is another feature of the invention to provide a method for producing a genetic map for a species, including: a) determining a pattern of hybridization products on an array for sets of samples, each sample within a set containing a different collection of fractionated genomic nucleic acid from a member of the species, the member is different for each set, the array includes a plurality of nucleic acid molecules, each nucleic acid molecule includes a nucleic acid sequence corresponding to a different sequence within the genome of the species, and the hybridization products are formed between the nucleic acid molecules and the fractionated genomic nucleic acid, and b) determining the relationship between nucleic acid sequences within the genome based on the pattern of hybridization products for each sample of each set and the genetic relationship of the different members for each set, thereby forming the genetic map.

It is a feature of the above-described method that the sets contain at least five,

or at least ten sets. Each set may contain at least five, or at least ten samples. In one aspect of the invention, the genomic nucleic acid may be digested with at least two, or at least five restriction enzymes. In addition, the fractionated genomic nucleic acid may be labeled. It is a further feature of the invention that each nucleic acid molecule is unique. The array may contain at least about 100, at least about 500, or at least about 1000 nucleic acid molecules. It is an intention of the invention that every twenty-five cM region, or, for instance, every two cM region of the genome contains at least one of the nucleic acid sequences. As used above, determining the relationship between each the nucleic acid sequence within the genome can be determining the relative position of each the nucleic acid sequence within the genome, or determining the relative distance between each of the nucleic acid sequences within the genome.

It is yet another feature of the invention to provide a method of producing a genetic map for a species, the method including contacting an array with sets of samples, wherein each sample within a set contains a different collection of fractionated genomic nucleic acid from at least one member of the species, the member(s) being different for each set, wherein the array comprises a plurality of nucleic acid molecules, wherein each nucleic acid molecule comprises a nucleic acid sequence corresponding to a different sequence within the genome of the species. The contacting is performed such that a pattern of hybridization products is formed for each sample of each set, the hybridization products being formed between the nucleic acid molecules and the fractionated genomic nucleic acid, wherein the relationship between the nucleic acid sequences within the genome is determinable based on the pattern of hybridization products for each sample of each set and the genetic relationship of the different members for each set. The relationship constitutes the genetic map.

In another aspect, the invention provides a method for identifying a region of a genome of a species, the region containing a nucleic acid sequence that contributes to a phenotype observed in at least one member of the species, the method including: a) determining a first group of patterns of hybridization products on an array for samples of a first set, wherein each sample within the first set comprises a different collection of fractionated genomic nucleic acid from the member(s). The array contains a

plurality of nucleic acid molecules, with each nucleic acid molecule having a nucleotide sequence corresponding to a different sequence within the genome of the species, wherein hybridization products are formed between the nucleic acid molecules and the fractionated genomic nucleic acid, b) determining at least one second group of patterns of hybridization products on the array for samples of at least one second set, wherein each sample within the second set comprises a different collection of fractionated genomic nucleic acid from at least one second member, the second member(s) being different for each second set, and c) identifying the region based on a comparison between the first and second groups of patterns of hybridization products and the genetic relationship between the member(s) and each second member(s). A representative species is maize. Further, a representative phenotype is a growth characteristic.

In another aspect of the invention, there is provided a method for identifying a region of a genome of a species, the region containing a nucleic acid sequence that contributes to a phenotype observed in a member of the species. The method includes contacting an array with a first set of samples and at least one second set of samples, each sample within the first set containing a different collection of fractionated genomic nucleic acid from the member, wherein each sample within the second set contains a different collection of fractionated genomic nucleic acid from a second member, the second member being different for each second set, wherein the array contains a plurality of nucleic acid molecules, wherein each nucleic acid molecule has a nucleic acid sequence corresponding to a different sequence within the genome. The contacting is performed such that a first group of patterns of hybridization products is formed for each sample of the first set and a second group of patterns of hybridization products is formed for each sample of the second set, the hybridization products being formed between the nucleic acid molecules and the fractionated genomic nucleic acid. The region is identifiable based on a comparison between the first and second groups of patterns of hybridization products and the genetic relationship between the member and each second member.

Another feature of the invention is a method of genotyping a member of a species, the method including determining a pattern of hybridization products on an array for a plurality of samples, wherein each sample contains a different collection of

fractionated genomic nucleic acid from the member, wherein the array contains a plurality of nucleic acid molecules, wherein each nucleic acid molecule has a nucleotide sequence corresponding to a different sequence within the genome of the species, wherein the hybridization products are formed between the nucleic acid molecules and the fractionated genomic nucleic acid, wherein the pattern indicates the genotype of the member.

In yet another feature, the invention provides a method of genotyping a member of a species, the method comprising contacting an array with a plurality of samples, wherein each sample contains a different collection of fractionated genomic nucleic acid from the member, wherein the array contains a plurality of nucleic acid molecules, wherein each nucleic acid molecule has a nucleic acid sequence corresponding to a different sequence within the genome of the species, wherein the contacting is performed such that a pattern of hybridization products is formed for each sample, the hybridization products being formed between the molecules and the fractionated genomic nucleic acid, wherein the pattern for each sample indicates the genotype of the member.

The invention further provides a method of genotyping a nucleic acid sample, the method comprising determining a pattern of hybridization products on an array for a plurality of fractions, wherein each fraction contains a different collection of fractionated genomic nucleic acid from the nucleic acid sample, wherein the array contains a plurality of nucleic acid molecules, wherein each nucleic acid molecule has a nucleotide sequence corresponding to a different sequence within a genome of a species, wherein the hybridization products are formed between the nucleic acid molecules and the fractionated genomic nucleic acid, wherein the pattern for each fraction indicates the genotype of the nucleic acid sample.

Additionally provided by the invention is a method of genotyping a nucleic acid sample. The method includes contacting an array with a plurality of fractions, wherein each fraction contains a different collection of fractionated genomic nucleic acid from the nucleic acid sample, wherein the array contains a plurality of nucleic acid molecules, wherein each nucleic acid molecule has a nucleic acid sequence corresponding to a different sequence within a genome of a species, wherein the contacting is performed such that a pattern of hybridization products is formed for

each fraction, the hybridization products being formed between the nucleic acid molecules and the fractionated genomic nucleic acid, wherein the pattern for each fraction indicates the genotype of the nucleic acid sample. For example, the nucleic acid sample may include genomic nucleic acid from a member of the species or from more than one member of the species. A representative nucleic acid sample is from a blood sample.

In yet another aspect of the invention, there is provided a method of producing a genetic map for a species, comprising performing amplification reactions on a plurality of samples using a plurality of IDP primer pairs, wherein each sample contains genomic nucleic acid from a different member of the species, wherein each IDP primer pair amplifies a different nucleic acid region within the genome of the species, wherein each nucleic acid region contains a different IDP, wherein the amplification reactions are performed such that the presence or absence of each different IDP is determined for each sample, and wherein the relationship between each different nucleic acid region within the genome is determinable based on the presence or absence of each different IDP and the genetic relationship of the different members. The relationship constitutes the genetic map. For example, the species may be a plant species, which may be maize. It is a feature of the invention that the plurality of samples contains at least five or at least ten samples. The plurality of IDP primer pairs may have at least about 500, or at least about 1000 IDP primer pairs. It is advantageous that every twenty-five cM, for example, every two cM region of the genome contain at least one of the nucleic acid regions. As used above, determining the relationship between each nucleic acid region within the genome can be used to determine the relative position of each nucleic acid region within the genome, or the relative distance between each nucleic acid region within the genome.

The invention further features a method for identifying a region of a genome of a species, the region containing a nucleic acid sequence that contributes to a phenotype observed in at least one member of the species. The method includes: a) performing a first set of amplification reactions with a sample containing genomic nucleic acid from the member(s) and a plurality of IDP primer pairs, with each IDP primer pair amplifying a different nucleic acid region within the genome of the species, wherein each nucleic acid region contains a different IDP, wherein the first

set of amplification reactions is performed such that the presence or absence of each different IDP is determined for the member(s), and b) performing a subsequent set of amplification reactions with at least one subsequent sample and the plurality of IDP primer pairs, wherein each subsequent sample contains genomic nucleic acid from at least one subsequent member of the species, the subsequent member(s) being different for each subsequent sample, wherein the subsequent set of amplification reactions is performed such that the presence or absence of each different IDP is determined for the subsequent member(s), the region being identifiable based on a comparison between the results of the first and subsequent sets of amplification reactions and the genetic relationship between the member(s) and each subsequent member(s).

The invention also features a method of genotyping a member of a species, the method comprising performing a set of amplification reactions with a sample containing genomic nucleic acid from the member and a plurality of IDP primer pairs, wherein each IDP primer pair amplifies a different nucleic acid region within the genome of the species, wherein each nucleic acid region contains a different IDP, wherein the set of amplification reactions are performed such that the presence or absence of each IDP is determinable for the member. The presence or absence of each IDP indicates the genotype of the member.

In addition, the invention features a method of genotyping a nucleic acid sample, the method comprising performing a set of amplification reactions with the nucleic acid sample and a plurality of IDP primer pairs, wherein each IDP primer pair amplifies a different nucleic acid region within a genome of a species, wherein each nucleic acid region contains a different IDP, wherein the set of amplification reactions is performed such that the presence or absence of each IDP is determinable for the nucleic acid sample, wherein the presence or absence of each IDP indicates the genotype of the nucleic acid sample. The nucleic acid sample may contain genomic nucleic acid from one or more members of the species.

Another feature of the invention is a genotyping method. The method includes contacting an array with a plurality of samples to form a pattern of hybridization products for each sample, each sample containing a different collection of fractionated genomic nucleic acid. The fractioned genomic nucleic acid can be labeled.

An additional feature of the invention is a method for identifying a nucleic

acid sequence that is regulated by a second nucleic acid sequence. The method includes, a) determining a first pattern of hybridization product intensities on an array, wherein the array contains a plurality of nucleic acid molecules, wherein each nucleic acid molecule has a nucleotide sequence corresponding to a different sequence transcribed by a member of a species, the first pattern of hybridization product intensities being formed between a first pool of nucleic acid and the nucleic acid molecules, wherein the first pool of nucleic acid corresponds to mRNA and is obtained from a first group of individuals from the species, wherein the first group of individuals have the second nucleic acid sequence, and b) determining a second pattern of hybridization product intensities on the array, the second pattern of hybridization product intensities being formed between a second pool of nucleic acid and the nucleic acid molecules, wherein the second pool of nucleic acid corresponds to mRNA and is obtained from a second group of individuals from the species, wherein the nucleic acid sequence is identifiable based on a comparison between the first and second patterns of hybridization product intensities. In one aspect, the first and second groups of individuals are progeny of the same parental cross. In addition, the first pool of nucleic acid may be mRNA, and further may be labeled. By way of example, the second pool of nucleic acid may be mRNA and also may be labeled. The nucleic acid molecules can be expressed sequence tags from the species.

In another aspect of the invention, there is provided a method for identifying a nucleic acid sequence that is regulated by a second nucleic acid sequence, the method comprising contacting an array with first and second pools of nucleic acid, wherein the array contains a plurality of nucleic acid molecules, wherein each nucleic acid molecule has a nucleotide sequence corresponding to a different sequence transcribed by a member of a species, wherein the first pool of nucleic acid corresponds to mRNA and is obtained from a first group of individuals from the species, wherein the first group of individuals have the second nucleic acid sequence, wherein the second pool of nucleic acid corresponds to mRNA and is obtained from a second group of individuals from the species, wherein the second group of individuals do not have the second nucleic acid sequence, wherein the contacting is performed such that a first pattern of hybridization product intensities is formed between the first pool of nucleic acid and the nucleic acid molecules and a second pattern of hybridization product

intensities is formed between the second pool of nucleic acid and the nucleic acid molecules. The nucleic acid sequence is identifiable based on a comparison between the first and second patterns of hybridization product intensities.

In an additional aspect of the invention, a method for detecting a polymorphism in a member of a species is provided, the method comprising: a) performing an amplification reaction with genomic nucleic acid from the member and a primer pair such that a product is formed if the genomic nucleic acid contains the polymorphism, and b) detecting the presence or absence of the product without size-fractionation. By way of example, the polymorphism may be an IDP, and the primer pair an IDP primer pair. In addition, for purposes of detection, the amplification reaction may contain a molecule for detection of the product, which may be ethidium bromide.

In general, the invention features a method for obtaining a primer pair that detects an IDP marker. The method includes a) obtaining a first sequence of a first DNA fragment, where the first DNA fragment is from a first allele; b) obtaining a second sequence of a second DNA fragment, where the second DNA fragment is from a second allele; c) selecting a first primer sequence that both the first and second DNA fragments contain; and d) selecting a second primer sequence that only one of the first and second DNA fragments contain. The first and second primer sequences are a primer pair that detects an IDP marker. The alleles can be from maize. The first and second DNA fragments can contain an RFLP marker.

In another aspect, the invention features a method for detecting a polymorphism (e.g., IDP) in an organism. The method includes a) obtaining genomic DNA from the organism; b) obtaining a first and second primer, where the first primer corresponds to an inserted or substituted DNA sequence of the polymorphism; c) performing an amplification reaction with the genomic DNA and the first and second primers such that a product is formed if the genomic DNA contains the inserted or substituted DNA sequence; and d) detecting the presence or absence of the product without size-fraction. The amplification reaction can contain an intercalating molecule (e.g., ethidium bromide).

Another aspect of the invention features a mapping array having a nucleic acid component consisting essentially of non-redundant nucleic acid fragments (e.g., more

than 500, 1000, 2000, 5000, or 10,000 non-redundant nucleic acid fragments). In another embodiment, the invention features an isolated collection of more than 500 (e.g., more than 750, 1000, 1500, 2000, 3000, 4000, 5000, 7500, or 10,000) nucleic acid fragments consisting essentially of non-redundant nucleic acid fragments.

Another aspect of the invention features a method for determining the genotype of a member of a species. The method includes a) obtaining an array having a plurality of DNA fragments; b) contacting the array with a series of labeled genomic DNA fractions from the member to form hybridization products between the labeled genomic DNA fractions and the DNA fragments; and c) determining the pattern of the hybridization products on the array. The pattern indicates the genotype. The array can have a nucleic acid component consisting essentially of non-redundant nucleic acid fragments (e.g., more than 500, 1000, 2000, 5000, or 10,000 non-redundant nucleic acid fragments).

Unless otherwise defined, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. Although methods and materials similar or equivalent to those described herein can be used in the practice or testing of the present invention, suitable methods and materials are described below. All publications, patent applications, patents, and other references mentioned herein are incorporated by reference in their entirety. In case of conflict, the present specification, including definitions, will control. In addition, the materials, methods, and examples are illustrative only and not intended to be limiting.

Other features and advantages of the invention will be apparent from the following detailed description, and from the claims.

DESCRIPTION OF DRAWINGS

Figure 1 contains a sequence alignment of intron 3 from B73 and Mo17 alleles of the *a1* gene. Intronic sequences are depicted in bold red, while flanking exonic sequences are blue.

Figure 2 is a diagram depicting the *umc102* alleles from the LH82 and GLAS maize lines as well as the relative position of the 102-L, 102-G, and 102-R primers.

Figure 3 contains photographs of an electrophoresis gel and solid support

containing corresponding IDP droplets.

Figure 4 is a diagram depicting the relative position of P1, P2, P3, and P4 primers.

Figure 5 contains a diagram depicting the relative position of the GS and PTN primers as well as photographs of electrophoresis gels stained with ethidium bromide (EtBr) (left) or probed with a labeled gel slice (right).

Figure 6 contains a photograph of a Southern blot identifying an RFLP.

Figure 7 contains a photograph of an electrophoresis gels containing size-fractioned genomic DNA stained with EtBr.

Figure 8 contains three photographs of electrophoresis gels treated as indicated.

DETAILED DESCRIPTION

The invention provides methods and materials related to the analysis of a genome. Specifically, the invention provides arrays, collections of IDP primer pairs, methods for producing a genetic map of a species, methods for genotyping, methods for identifying nucleic acid sequences that regulate another sequence, and methods for identifying nucleic acid sequences that are regulated by another sequence.

Arrays

The invention provides various arrays that can be used to analyze a genome. The term "array" as used herein refers to a collection of nucleic acid molecules that are arranged in defined areas such that each defined area contains at least one copy of a particular nucleic acid molecule. For example, an array can have a collection of nucleic acid molecules on a glass slide arranged in a series of spots organized into multiple rows and columns. Typically, each defined area contains many copies of the same nucleic acid molecule. The collection of nucleic acid molecules of an array can be redundant or non-redundant. In addition, the sequence of each nucleic acid molecule of an array can be known, partially known, or unknown. Each array of the invention contains a nucleic acid component that can be attached to any solid support such as those described in U.S. Patent Number 6,040,193. For example, an array can have a collection of nucleic acid molecules deposited on a slide or chip at a particular

density. Other examples of solid supports include, without limitation, glass, Pyrex, quartz, silicon, polystyrene, and polycarbonate. Any method can be used to make an array such as those described elsewhere (e.g., U.S. Patent Numbers 6,040,193; 6,054,270; and 5,800,992). For example, the nucleic acid component of an array can be deposited on a solid support using spotting techniques (e.g., spotting via a robotic system), channel flow technology, attachment to linker molecules, light-directed synthesis techniques (e.g., deprotection and coupling using a binary mask), and computer-controlled printing device technology (e.g., pen plotter).

In one embodiment, the invention provides an array having a nucleic acid component consisting essentially of non-redundant nucleic acid molecules. The term "nucleic acid component" as used herein with respect to an array refers to the entire portion of the array that is made of nucleic acid. Thus, each array has a single nucleic acid component. The term "non-redundant" as used herein with respect to nucleic acid molecules of different defined areas means that the sequence of the nucleic acid molecules in one defined area is different from the sequence of the nucleic acid molecules of the other defined areas of the array. For example, a collection of nucleic acid molecules of an array would be considered completely non-redundant if no two nucleic acid molecules from different defined areas of that array were identical. Likewise, a collection of nucleic acid molecules of an array would be considered highly redundant if the nucleic acid molecule in each defined area of the array was present in more than one defined area. It will be appreciated that an array having a nucleic acid component consisting essentially of non-redundant nucleic acid molecules can contain a limited number of defined areas each containing the same nucleic acid molecule. Thus, a nucleic acid component of an array would be considered to consist essentially of non-redundant nucleic acid molecules even though the same nucleic acid molecule was represented a few times in different defined areas. For example, the same nucleic acid molecule can be located in more than one defined area of an array to serve as a control. Furthermore, a single solid support may contain one array or multiple arrays. If a solid support contains more than one array, the array may be different arrays (i.e., different nucleic acid components) or may be the same array duplicated on the support.

For the purposes of this invention, the term "nucleic acid" encompasses both

RNA and DNA, including cDNA, genomic DNA, and synthetic (e.g., chemically synthesized) DNA. The nucleic acid can be double-stranded or single-stranded. Where single-stranded, the nucleic acid can be the sense strand or the anti-sense strand. In addition, nucleic acids can be circular or linear.

An array can contain any type of nucleic acid from any source. For example, an array can contain, without limitation, DNA, cDNA, genomic DNA, mRNA, chloroplast DNA, mitochondria DNA, or combinations thereof. In addition, an array can contain synthetic nucleic acid or nucleic acid corresponding to nucleic acid from an organism. For example, a nucleic acid molecule of an array can contain a nucleic acid sequence corresponding to a sequence from any organism including, without limitation, plants (e.g., corn, wheat, rice, tobacco, cotton, sunflower, and vegetable plants), animals (e.g., humans, cows, sheep, chickens, pigs, dogs, and fish), and microorganisms (e.g., bacteria, fungus, and algae). In some cases, the nucleic acid sequence can correspond to a sequence from a virus (e.g., retroviruses, reoviruses, herpesviruses, and influenza viruses). When a sequence corresponds to a sequence of an organism, that sequence can be a genomic sequence, a transcribed sequence, or a transcribed and translated sequence.

At least about 50 percent of the non-redundant nucleic acid molecules of an array of the invention can have a nucleic acid sequence corresponding to an untranslated sequence in an organism. The term "untranslated sequence" as used herein refers to those nucleic acid sequences that may or may not be transcribed, but are not translated. For example, sequences that are typically transcribed, but are untranslated, can be a 5' untranslated region (5' UTR), a 3' untranslated region (3' UTR), or an intronic sequence. Untranslated sequences can be identified from a genomic DNA, cDNA, or mRNA sequence by eye or through the use of computer software designed to locate, for example, start codons, mRNA splice sites, coding sequences, stop codons, and polyadenylation sites. Alternatively, at least about 75 percent, or at least about 90 percent, or at least about 95 percent of the non-redundant nucleic acid molecules of an array of the invention can have a nucleic acid sequence corresponding to an untranslated sequence in an organism. In addition, non-redundant nucleic acid molecules of an array of the invention may include non-transcribed sequences, such as promoter regions or intergenic (e.g., non-genic) regions.

The nucleic acid component of an array can contain nucleic acid molecules that lack repeated sequences. The term "repeated sequences" as used herein refers to nucleic acid sequences that are (1) at least about 30 nucleotides in length, (2) identical or nearly identical (i.e., greater than 90 percent identity) to each other, and (3) present in a genome more frequently than would be statistically expected based on the length of the sequence, the identity, and the size of the genome. Repeated sequences include, without limitation, transposable elements and microsatellites.

An array can contain any number of nucleic acid molecules at any density. Typically, an array of the invention contains more than about 500 nucleic acid molecules (e.g., more than about 750, 1000, 1500, 2000, 2500, 5000, 10000, 15000 nucleic acid molecules) at a density of about 100 or more (e.g., about 250, 500, 1000, 2000, 5000, or more) defined areas per square centimeter. In one embodiment, an array can contain a collection of nucleic acid molecules having sequences corresponding to sequences in a genome such that at least every fifty cM region (e.g., at least every 25, 20, 15, 10, 5, 2, 1, or 0.5 cM region) of the genome contains at least one of the corresponding sequences.

The nucleic acid component of an array can have redundant or non-redundant nucleic acid molecules. In addition, the nucleic acid component of an array can contain one or more groups of nucleic acid molecules (e.g., two, five, ten, twenty, or more groups). Typically, each nucleic acid molecule within a group has a nucleic acid sequence corresponding to a sequence that is transcribed by a cell from a particular source. For each group, the source can be different. For example, one group of nucleic acid molecules of a nucleic acid component can have sequences corresponding to sequences transcribed by a cell from root tissue of a corn plant, while a second group of nucleic acid molecules of the nucleic acid component can have sequences corresponding to sequences transcribed by a cell from stem tissue of a corn plant. The source can be any source such as tissue at a particular stage of development. For animals, the source can be, without limitation, organ tissue (e.g., liver, brain, skin, heart, lung, or kidney) and cellular samples (e.g., white blood cells, tumors, or nerves) at any stage of development (e.g., embryonic, birth, yearling, or adult). For plants, the source can be, without limitation, organ tissue such as roots, shoots, stems, leaves, flowers, or such organ tissue or seeds and plants at any stage of development (e.g.,

seedlings or full grown plants), or may be from, for example, in inbred line, a hybrid, or a plant carrying a mutation. In addition, the nucleic acid component can be obtained from a particular source as outlined above following exposure to one or more conditions (e.g., drought, cold, salt, light, or disease).

Each nucleic acid molecule within a group can contain a marker such that the source of that nucleic acid molecule can be identified. For example, each nucleic acid molecule from the root of a corn plant can have a nucleic acid marker having a specific sequence that identifies those nucleic acid molecules as being from the root of a corn plant. Nucleic acid molecules having such markers can be made using any method. For example, mRNA isolated from the root of a corn plant can be used to make cDNA in a manner such that a linker sequence containing a marker is added to one of the ends of each newly synthesized cDNA. Thus, every cDNA made from the mRNA isolated from a corn plant root will have the same identifiable marker. A marker can be of any type. For example, nucleic acid, chemical, or radioactive markers can be used. A nucleic acid marker can be any length (e.g., about 10, 15, 20, 25, or 30 nucleotides) and can have any sequence provided that it can be used to identify the source of a nucleic acid molecule. It will be appreciated that the presence of the same nucleic acid marker in otherwise different nucleic acid molecules within a group does not change a non-redundant collection into a redundant collection.

In one embodiment, an array can have a nucleic acid component that has ten groups of nucleic acid molecules. Each group can have nucleic acid molecules with sequences corresponding to sequences transcribed by cells from different tissue of a corn plant. For example, one group can contain nucleic acid molecules corresponding to sequences transcribed by root cells, while another group contains nucleic acid molecules corresponding to sequences transcribed by stem cells, and yet another group contains nucleic acid molecules corresponding to sequences transcribed by leaf cells. A marker specific for each group can be incorporated into each nucleic acid molecule of a group. Any method can be used to make the various groups of nucleic acid molecules. For example, standard library construction techniques (e.g., cDNA or genomic DNA library construction techniques) can be used to make large groups of nucleic acid molecules. In addition, chemical synthesis techniques can be used to make large groups of nucleic acid molecules. The nucleic acid molecules of one

group can be made separately from the nucleic acid molecules of another group. Once made, the nucleic acid molecules of each group can be pooled. The nucleic acid molecules between groups can be redundant or non-redundant. If desired, any redundant nucleic acid molecules between groups can be removed using any method. For example, varying degrees of subtractive hybridization techniques can be used to make a redundant collection less redundant.

An array can be used in a hybridization reaction once or more than once. Thus, it will be appreciated that the descriptions used herein that refer to contacting multiple samples to "an array" means that either (1) the exact same physical array is re-used for each sample, or (2) a different physical array from a supply of identical arrays is used for each sample.

IDP primer pairs

The invention also provides IDP primer pair collections. An IDP is an insertion/deletion polymorphism. The term "IDP primer pair" as used herein refers to a pair of primers that can amplify nucleic acid containing an IDP selectively by having one primer that hybridizes to a nucleic acid sequence common among different alleles and another primer that hybridizes to a nucleic acid sequence containing an IDP. When a sample contains nucleic acid having the particular IDP recognized by the IDP primer pair, then a detectable amplification product will be produced. This amplification product can be detected using any method including, without limitation, visual and size-fractionation techniques. For example, ethidium bromide can be added to the amplification reaction mixture during or after completion of the amplification reaction such that the accumulation of an amplification product can be detected visually without size-fractionation (e.g., gel electrophoresis, HPLC, or the like). Typically, the sequence of each primer of an IDP primer pair is known. In some cases, however, the sequence of some or all of the primers can be unknown. In addition, primers may be degenerate or may be a combination of primer sequences (e.g., hexamers).

Any method can be used to identify IDPs, such as sequencing or denaturing HPLC (dHPLC). For example, sequence alignments between two alleles can be used to locate IDPs. Typically, untranslated sequences within the genome of a species

contain more IDPs than translated sequences. Thus, sequencing efforts can be focused on untranslated regions of different alleles such that IDPs are readily identified. In addition, the amount of sequencing necessary to identify an IDP can be reduced by first locating an untranslated region within a database (e.g., GenBank) and then sequencing the same untranslated region from a different allele.

Once an IDP has been identified, any method can be used to design an IDP primer pair specific for that IDP. For example, the sequence for each primer of an IDP primer pair can be designed by hand using a sequence alignment between two alleles. In addition, a computer can be used to design IDP primer pairs based on a set of predetermined parameters such as the length of each primer, the length to be amplified, nucleotide content, and the like. It will be appreciated that at least two IDP primer pairs can be designed for each IDP. One IDP primer pair can be designed to recognize the IDP of one allele, while another IDP primer pair can be designed to recognize the IDP of another allele. In the situation, the first primer of each IDP primer pair can be identical, while the second is specific for the IDP of each allele.

An IDP primer pair collection can contain any number of IDP primer pairs. For example, an IDP primer pair collection can contain 100, 250, 500, 1000, 2500, 5000, 10000, or more IDP primer pairs. In one embodiment, an IDP primer pair collection can be such that at least every fifty cM region (e.g., at least every 25, 20, 15, 10, 5, 2, 1, or 0.5 cM region) of the genome of a species contains at least one nucleic acid segment targeted by an IDP primer pair in the collection.

Genetic maps, identifying genes, and genotyping

The invention provides methods for producing a genetic map of any species (e.g., plants, animals, or microorganisms). The term "genetic map" as used herein refers to the arrangement of nucleic acid sequences within the genome of a species. Genetic maps can have various levels of detail. For example, a genetic map can be such that the arrangement of every nucleic acid sequence of a genome is known, or a genetic map can be such that the arrangement of some portion less than all the nucleic acid sequences of a genome is known.

The invention provides the following methods for making a genetic map. First, different members of a species or members of two distinct species that are inter-

fertile are selected. Any number of members can be selected. It is noted that the analysis of a larger number of members provides more information than the analysis of a smaller number of members. Typically, the genetic relationship between each selected member is known. Once selected, a genomic nucleic acid sample is collected from each member. Any method can be used to collect genomic nucleic acid. Once collected, it is desired that the genomic nucleic acid be fractionated. Any method can be used to fractionate the genomic nucleic acid, for example, size fractionation, or fractionation based on GC content or methylation state. For instance, to fractionate the genomic nucleic acid based upon size, the genomic nucleic acid can be digested with one or more restriction enzymes (e.g., two, three, four, five, six, or more restriction enzymes, alone or in various combinations). Any type of restriction enzyme can be used. For example, frequent cutters or infrequent cutters can be used. Once digested, the genomic nucleic acid from each member can be divided into a series of fractions based on size. For example, the digested genomic nucleic acid can be separated by gel electrophoresis and divided into multiple samples by cutting the gel into gel slices such that each gel slice contains genomic nucleic acid of a particular size range. The digested genomic nucleic acid can be divided into any number of fractions (e.g., 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, or more fractions). In addition, each fraction can contain any size range. At this point, a set of fractionated genomic nucleic acid samples results for each member selected. For example, if five members were selected, then five sets of fractionated genomic nucleic acid samples are produced. In addition, if the genomic nucleic acid from all five members was fractionated into ten samples, then five sets with each set containing ten fractionated genomic nucleic acid samples are produced. Thus, each set contains a series of fractionated genomic nucleic acid samples from a particular member of a species. It is noted that the size-parameters for each fraction within a set should be the same for each set being compared to one another. In addition, the fractionated genomic nucleic acid can be labeled. For example, the fractionated genomic nucleic acid of each sample for each set can be radioactively labeled.

Once the sets of fractionated genomic nucleic acid samples are obtained, each sample from each set is contacted with an array such that a pattern of hybridization products is formed for each sample from each set. As described herein, an array

contains a collection of nucleic acid molecules. Since each nucleic acid molecule on the array that has a sequence corresponding to a sequence within the genome of the selected members can be genetically mapped, the array used in such mapping methods typically contains a large collection of nucleic acid molecule known to have sequences corresponding to the sequences within the genome of the selected members. For example, an array can contain fragments of nucleic acid from the same species as that of the selected members. In addition, the array can have any of the properties described herein. The hybridization products are formed between any nucleic acid molecule of the array and any fractionated genomic nucleic acid that have corresponding sequences.

Once a pattern of hybridization products is obtained for each sample from each set, the relationship (e.g., relative order or relative distance) between each nucleic acid molecule on the array that has a sequence corresponding to a sequence within the genome of the selected members can be determined based on the pattern of hybridization products for each sample of each set and the genetic relationship between each selected member. A computer can be used to analyze the patterns for each sample of each set and the genetic relationship between the selected members such that the nucleic acid sequence on the array are arranged into a genetic map. Thus, determining the pattern of hybridization products that are produced on an array for each of a series of fractionated genomic nucleic acid samples from different members of the species whose genome is to be mapped, and then determining the relationship between each nucleic acid molecule on the array that has a sequence corresponding to a sequence within the genome of the selected members based on (1) the pattern of hybridization products for each sample of each set and (2) the genetic relationship between each selected member can be used to arrange a large number of nucleic acid sequences of a genome into a genetic map.

The fractionated genomic nucleic acid samples and arrays described herein also can be used to identify regions of a genome responsible for any phenotype based on (1) a comparison of the patterns of hybridization products on an array for each fractionated genomic nucleic acid sample from each member of a group of members from a species, (2) the genetic relationship between each member, and (3) the presence or absence of the particular phenotype being analyzed in each member. It

should be appreciated that regions of a genome responsible for a phenotype may be polymorphic relative to the group of members (e.g., member(s) may possess an insertion, substitution or deletion relative to other members of the group), or the phenotype may be due to differences in the level of a gene's expression within the group of members.

In addition, the fractionated genomic nucleic acid samples and arrays described herein can be used in genotyping. For example, any genomic nucleic acid sample can be isolated, digested, and fractionated to produce a series of fractionated genomic nucleic acid samples that can be analyzed on an array to produce a pattern of hybridization products for each sample. The patterns of each sample reflect the genotype for that particular sample. The genomic nucleic acid sample can be genomic nucleic acid from a single individual or genomic nucleic acid from a population of individuals. The individual can be from the same species or different species. The genotyping methods and materials described herein can be used in marker-assisted breeding, forensics, identification and tracking of inbred line or germplasm and paternity and maternity determinations.

The invention also provides a method for producing a genetic map that involves performing amplification reactions on multiple genomic nucleic acid samples using one of the collections of IDP primer pairs described herein such that the presence or absence of each IDP recognized by each IDP primer pair is determined for each sample. Typically, each genomic nucleic acid sample is from a different member of the species whose genome is to be mapped. It is noted that the analysis of a larger number of samples provides more information than the analysis of a smaller number of samples. Once the amplification reactions are performed, the relationship between each nucleic acid region containing each IDP within the genome can be determined based on the presence or absence of each IDP recognized by each IDP primer pair and the genetic relationship of the different members from which the samples were collected. Again, a computer can be used to analyze this information and arrange the nucleic acid regions amplified by the IDP primer pairs into a genetic map. It will be appreciated that a genetic map can be produced using a combination of methods.

The collections of IDP primer pairs described herein also can be used to identify regions of a genome responsible for any phenotype based on (1) a comparison

of the presence or absence of each IDP recognized the IDP primer pairs for a group of members of a species, (2) the genetic relationship between each member, and (3) the presence or absence of the particular phenotype being analyzed in each member.

In addition, the collections of IDP primer pairs described herein can be used in genotyping. For example, any nucleic acid sample can be analyzed using a collection of IDP primer pairs to determine the presence or absence of each IDP recognized by the IDP primer pairs in the nucleic acid sample. The presence or absence of each IDP indicates the genotype of the nucleic acid sample. The nucleic acid sample can be nucleic acid from a single individual or nucleic acid from a population of individuals. The individual can be from the same species or different species. The genotyping methods and materials described herein can be used in marker-assisted breeding, forensics, identification and tracking of inbred line or germplasm and paternity and maternity determinations.

Gene regulation

The invention provides methods for identifying a nucleic acid sequence that regulates another nucleic acid sequence within a genome as well as methods for identifying a nucleic acid sequence that is regulated by another nucleic acid sequence within a genome. An array containing nucleic acid molecules having nucleic acid sequences corresponding to transcribed sequences of a species can be contacted with two pools of nucleic acid corresponding to mRNA (e.g., mRNA and cDNA) to produce two patterns of hybridization product intensities. The first pool of nucleic acid corresponding to mRNA is from a group of individuals having a particular nucleic acid sequence, while the second pool is from a group of individuals having a different nucleic acid sequence that corresponds to the nucleic acid sequence from the first group of individuals. For example, the individuals of the first pool can have allele A at region #1, and the individuals of the second pool can have allele B at region #1. In this case, nucleic acid molecules on the array that produced significant hybridization product intensities for the first pool, but not the second pool, can be identified as being regulated by the nucleic acid sequence of allele A at region #1. In addition, the nucleic acid sequence of allele A at region #1 can be identified as being a sequence that regulates another sequence. It is noted that the individuals in each of

the two pools can all be from a single parental cross.

Maize and other aspects of the invention

During most of its history, maize has been cultivated as open-pollinated varieties that consisted of collections of heterogeneous genotypes. Early in this century, however, it was demonstrated that homogenous pure (i.e., inbred) lines could be extracted from these varieties following five to seven generations of inbreeding. Although the resulting inbred lines were often quite weak, they could be intercrossed to produce vigorous and uniform F_1 hybrids. Indeed, some, but not all, of the resulting F_1 hybrids produced larger seed yields than the open-pollinated varieties from which the corresponding inbred parents were derived. This phenomenon is termed heterosis. Because of the large amount of heterosis that can be obtained in selected maize lines, essentially all maize grown in the United States is from hybrid seed.

Because not all F_1 hybrids are superior, a central problem that has faced plant breeders is how to identify which pairs of inbreds should be used to generate hybrids. Currently, elite hybrids are identified by inbreeding in two relatively narrow genetic groups called heterotic pools and then making crosses between inbreds derived from these two heterotic pools. The identification of elite hybrids is dependent on data collected from replicated yield trials. Despite the fact that hybrids have been developed in this manner for nearly 70 years, relatively little is known about the genetic basis of quantitative traits and heterosis.

The maize populations used herein (available from Iowa State University) have been under intensive genetic selection for a half century using, for example, reciprocal recurrent selection techniques. Reciprocal recurrent selection (RRS) is a plant breeding procedure that allows for the improvement of the average yields of F_1 hybrids generated from individuals derived from two populations. By its nature, RRS emphasizes selection for heterotic response. Since 1949, RRS has been conducted at Iowa State University on two maize populations, BSSS and BSCB1. The BSSS and BSCB1 populations were developed in the 1940s by intercrossing 16 and 12 inbred lines, respectively. Since that time, 15 cycles of RRS have been conducted on these populations. Briefly, individual pairs of plants (or their inbred progeny) from each

population were simultaneously self-pollinated and crossed to generate F_1 hybrid seed that was yield-tested in replicated field trials. Based on the results of these yield trials, between 10 and 20 self-pollinated lines from each population were selected for several generations of random mating to generate subsequent populations (cycles). Over 15 cycles of RRS, the yields of the two populations themselves have not increased substantially. However, the yields of crosses (i.e., hybrids) between random plants derived from two populations, have increased almost 7 percent per cycle. Indeed, between Cycles 0 and 11, the average amount of mid-parent heterosis between lines derived from these two populations has increased from 25 to 76 percent. Because this RRS program has been successful in selecting for increased yield and heterosis, the 15 cycles of the BSSS and BSCB1 populations and the 28 inbred lines from which they were derived represent an outstanding resource for the molecular study of the quantitative genetics of yield and heterosis.

The BSSS population has made significant contributions to the hybrid seed corn industry and U.S. agriculture. Inbred lines developed from BSSS (B14, B37, B73, B84) were direct parents of 19 percent of the total hybrid seed used to plant the maize acreage in the U.S. in 1980 and 42.2 percent of the hybrid seed produced for use in 1980 traced their origins to these inbred. Isozyme marker studies indicate that BSSS-related germplasm is present in more than 60 percent of the hybrids sold commercially in the U.S.

The creation of a dense, high-resolution genetic map has been hampered by the lack of genetic resolution in the widely used, public-domain maize mapping populations. This is because these populations were produced with minimal opportunities for recombination. For example, the three most widely used populations were created by crossing pairs of inbred lines and then deriving mapping progeny by self-pollination directly from F_2 plants. Burr and Burr (*Trends Genet.*, 7:55-60 (1991)) describe recombinant inbreds of Tx303 x CO159 and T232 x CM37, while Gardiner *et al.* (*Genetics*, 134:917-30 (1993)) describe immortalized F_2 plants of Tx303 x CO159. In addition, these populations have small sample sizes of progeny (n=54 or less). Although these maps have served as a very useful and central resource for many basic and applied initiatives in the plant sciences, the intermated B73 x Mo17 (IBM) population (n=350) was developed to meet the needs for an enhanced

mapping population. This was done by intermating an F_2 population derived from the single cross of the inbreds B73 and Mo17 for several generations prior to the extraction of recombinant inbred (RI) lines. The genetic resolution in the resulting population was therefore enhanced because additional opportunities for recombination were provided during the intermating generations. The value of the IBM population for mapping studies is further enhanced by the fact that populations derived from the B73 x Mo17 cross have been widely used in the study of quantitative trait loci.

Genetic markers are essential for the study of many fundamental biological processes. For example, they are needed to conduct evolutionary, population, and quantitative genetic studies. They also can be used to link gene sequences to function, for example, by comparing the genetic map positions of cDNAs to those of genes responsible for mutant phenotypes (i.e., candidate gene cloning). Finally, genetic markers can be used to cross-link genetic, physical, and cytological maps.

Microsatellites, simple sequence length polymorphisms (SSLPs), and simple sequence repeats (SSRs) are useful genetic markers because they are (1) highly polymorphic, (2) usually codominant, and (3) do not require a hybridization step. There are currently a few hundred mapped maize SSRs some of which are available on the internet at http://www.agron.missouri.edu/ssr.html.

Efforts to understand the genetic basis of heterosis and quantitative traits in genetically broad-based populations have been hampered by an absence of cost-effective, high-throughput, allele-specific markers. For example, the single "allele" detected by an RFLP probe in a genetically broad-based population may in fact represent two or more alleles that share a common restriction pattern but that have different DNA sequences and may therefore be functionally distinct. In addition, these analyses are complicated by the fact that maize is a diploidized tetraploid, and it is therefore not always clear whether distinct RFLP patterns represent alleles or duplicated genes.

Although SSRs offer several significant advantages over previous generations of markers (e.g., RFLPs and RAPDs), they still suffer from two disadvantages that limit their usefulness for the characterization of quantitative traits and heterosis. First, because SSR genotyping requires an electrophoresis step (often using expensive equipment), SSRs are not readily amenable to the high-throughput analyses required

for large-scale genetic studies. Second, given the high mutation rate at SSR loci, a particular SSR allele could have arisen independently two or more times over evolutionary time. This potential lack of allele-specificity limits the usefulness of SSRs in population studies.

In contrast to SSRs that require electrophoresis, genetic markers that yield plus/minus signals have the potential to be scored via chips. One such class of markers is single-nucleotide polymorphisms (SNPs). As genetic markers, SNPs have the advantage of being much more plentiful than other markers (e.g., SSRs). As described herein, the invention provides an alternative source of allele-specific genetic markers suitable for high-throughput screening: a novel class of co-dominant, allele-specific, PCR-based markers called insertion/deletion polymorphisms (IDPs).

Although the molecular basis of heterosis is not known, it is likely that alterations in the patterns of gene expression between inbreds and their hybrid progeny play at least some role. A number of emerging high-throughput technologies are revolutionizing the means by which gene expression research can be conducted. For example, DNA-based arrays that detect the accumulation of transcripts from thousands of genes in a single hybridization experiment have recently been developed. There are two significant concerns about using plant cDNAs as the targets for array-type experiments. First, the genomes of many important crop plants have undergone polyploidization events during their evolution. For example, maize is a segmental allotetraploid. As a consequence, at least two copies of most coding regions are present in the maize genome. These paralogous genes (e.g., genes A-1 and A-2) have the potential to confound the analysis of array data because there is often enough DNA sequence similarity with the paralogous genes causing cross-hybridization. Hence, if Gene A-2, but not Gene A-1, is expressed under State 1, cross-hybridization has the potential to indicate erroneously that Gene A-1 is also expressed under State 1. Such erroneous results have the potential to complicate data analysis from arrays; for instance, the computational discovery of DNA motifs that control state-specific gene expression (e.g., promoter elements).

A second concern relates to the retrotransposons that are present at high copy numbers in both the intergenic regions of the maize genome and in introns. Because these elements are present in cDNA pools, there exists a serious possibility of

retrotransposon-based cross-hybridization between cDNA targets and cDNA probes generating spurious gene expression data in array-type experiments. This would occur, for example, if (1) Gene A (which is not expressed under State 1) is represented on an array by an EST clone that contains a retrotransposon X (which perhaps went unrecognized because it resides in that portion of the clone that was not sequenced), (2) retrotransposon X is also present in the introns of other genes (B, C, D etc.) that are expressed under State 1, and (3) some fraction of the introns from genes B, C, or D are not correctly spliced (perhaps in a state-specific manner) in the cDNA pool used as a probe to study gene expression in State 1. Under these circumstances, hybridization could be observed to Gene A, even though Gene A is not expressed under State 1.

As described herein, the invention provides methods and materials for the high-throughput genetic mapping of cDNA (e.g., EST) clones and mutants as well as the generation and mapping of a new class of allele-specific markers (IDPs) that are suitable for high-throughput analyses. These methods and materials will enhance the study of genome-wide patterns of meiotic recombination, chromosome structure, gene distribution, and population genetics. They also can be used to refine quantitative genetic theory, conduct marker assisted selection (MAS) programs, and construct the specific genotypes required for quantitative genetic studies of, for example, gene expression, gene action, and gene interactions. In addition, the methods and materials can be used to link gene sequences to function via, for example, the genetic mapping of genes responsible for mutant phenotypes, candidate gene cloning, and QTL mapping, as well as by facilitating double mutant analyses and suppressor/enhancer screens.

The genetic markers provided herein can be used to (1) cross-link genetic, physical, and cytological maps, (2) set the stage for the positional cloning of genes, (3) conduct evolutionary studies, and (4) serve as starting points for the genomic sequencing of maize.

Further, the methods and materials described herein can relate to a single population that is being used as a mapping resource by other genome projects such that the generated data can readily be combined with those projects. For example, genetic mapping experiments can be conducted in a single maize population that is

being used as a mapping resource by other maize genome projects. In addition, the resulting dense genetic maps can be used to test for microsynteny among homologous and orthologous chromosomal segments to provide important information regarding organization and evolution of the maize genome.

In one embodiment, the invention provides an array-based mapping procedure that can be used to map genetically a non-redundant set of about, for example, 10,000 sequence-defined nucleic acid fragments, such as EST clones. It is important to note that EST clones and other cDNAs are used herein as examples, and other types of nucleic acid fragments such as synthetic nucleic acid molecules, genomic fragments, plasmid DNA, and viral nucleic acid can be used. Once the EST clones are mapped, they can be used as RFLP markers, and can facilitate candidate gene cloning efforts. For example, as groups of genes responsible for complex traits (e.g., yield and heterosis) are genetically mapped via QTL analyses, the methods and materials described herein can allow predictions to be made regarding which cDNAs are responsible for these traits. The mapping array also can be adapted to position genes responsible for simply inherited mutant phenotypes relative to the large collection (e.g., about 10,000) of mapped ESTs. In so doing, it will provide a tool for determining the functions of genes defined only by DNA sequence. In addition, the availability of these mapped EST clones can enhance existing genome research projects focused on developing physical and cytological maps of, for example, the maize genome. Further, given the species-independent nature of the high-throughput mapping methods and materials described herein, mapping arrays will have wide applicability in plant, animal, and human genomic research.

The invention provides co-dominant allele-specific markers (IDPs) for organisms such as maize as well as maps containing these markers. IDPs are PCR-based markers that detect the small insertions and deletions that occur at high frequencies among, for example, maize alleles. Like the allele-specific single nucleotide polymorphisms (SNPs) being developed as part of the Human Genome Project, IDPs are suitable for high-throughput analyses. Unlike SNPs, however, IDPs can be detected using a thermocycler and a UV light source. Hence, IDPs are suitable for use in most genetics laboratories including, without limitation, maize genetics laboratories. With respect to maize genetics, it is important to note that inbred lines

B73 and Mo17 as well as the BSSS and BSCB1 populations have been widely used by many of the world's breeding programs. Consequently, IDP markers identified from these lines are expected to occur at high frequencies in most commercially important breeding lines and populations. Thus, these IDP markers can have wide applicability in applied breeding efforts. Moreover, IDPs that detect the alleles from the parental inbreds of the BSSS and BSCB1 populations are extremely useful for population genetic studies in two of the world's best-studied maize populations. It also is important to note that polymorphisms detected by IDPs are unique enough that they are unlikely to have arisen independently. Thus, two alleles detected by an IDP marker are almost certainly related by decent; a feature that is not always true of SSRs or RFLPs alleles. Further, these populations have been subjected to a variety of selection schemes for agronomic traits. Thus, IDPs identified in these populations can be used to refine quantitative genetic theory. For example, using the high-throughput IDP genotyping methods and materials described herein, it will be possible to efficiently study genome-wide changes in allele frequencies that have occurred over many cycles of reciprocal recurrent selection (RSS) for heterosis in the BSSS and BSCB1 populations. In other words, the methods and materials of the invention can be used to define those chromosome segments that have undergone changes in frequency in these populations during selection for yield and heterosis. In addition, these methods and materials can be used to identify ESTs that reside in these chromosomal intervals as well as those genes whose expression is affected by these chromosomal intervals.

As described herein, gene expression studies can be conducted using arrays that facilitate the global analysis of mRNA levels. For example, the invention provides a collection of gene-specific "target" DNAs that can be spotted on a DNA chip. It is noted that using intact EST clones as "targets" can be problematic. First, chips using intact EST clones will often not be able to distinguish clearly between the expression patterns of sequence-related duplicate genes. Second, intact EST clones can contain unrecognized retrotransposons that have the potential to yield spurious expression data when used as targets on a DNA chip. As described herein, the methods and materials of the invention overcome these limitations by providing short sequence-defined 3'-UTR-enriched PCR products for use as targets on arrays. Thus,

the target sequences provided herein will have significant gene specificity and will not contain retrotransposons that can be recognized on the basis of sequence comparisons.

The invention will be further described in the following examples, which do not limit the scope of the invention described in the claims.

EXAMPLES

Example 1 - Isolation of IDP Markers

Analyses of the sequences of *a1* alleles from 24 maize lines revealed 11 haplotypes. The 1.2-kb region of the *a1* gene that was sequenced contained 23 nucleotide substitutions and 17 small insertion/deletions (indels) across the 11 haplotypes. In addition, a comparison of the sequences of intron 3 from the B73 and Mo17 alleles of the *a1* gene revealed the existence of at least four indels within the intron sequences (Figure 1). Thus, introns were found to be a particularly rich source of indels.

Example 2 - Converting RFLP markers into IDP markers

The DNA sequence of the umc102 plasmid that reveals an RFLP that maps to chromosome 3 was retrieved from GenBank. Based on this sequence, two primers were designed. These primers were used to amplify and sequence the *umc102* alleles from two maize lines (GLAS and LH82). Primers 102-G and 102-L were designed based on the indels revealed between these two alleles such that when used in combination with a non-specific primer (102-R), they yield PCR products only with GLAS and LH82 template DNAs, respectively (Figure 2).

Because IDP markers are scored on the basis of a plus/minus PCR assay, their detection does not require the time-consuming and often expensive electrophoresis step required for SSR detection. To detect the PCR product indicating the presence of one or more IDP markers, a 3-5 µL droplet of an IDP PCR reaction containing 1 µg/mL of EtBr was exposed to UV light (Figure 3; far right). Because EtBr is an intercalating dye, only PCR- positive droplets fluoresce (e.g., compare droplets 1 and 2). However, some IDP primer pairs routinely produce small amounts of heterogeneous, low-molecular weight, non-specific PCR products. For example, although not visible in the gel picture, such products were produced with primer pair

102-L/102-R (Figure 3; lanes 3 and 4). Presumably, this small amount of product was responsible for producing the small amount of fluorescence observed in droplet 3 (Figure 3). It is important to note that this non-specific fluorescence did not interfere with IDP scoring since it occurred equally across templates. Thus, scoring was straightforward if the florescence levels of experimental samples are directly compared to those of positive and negative controls of known genotypes. In particular, there was no difficulty in distinguishing the signals present in droplets 3 versus 4 (Figure 3). Thus, IDPs were detected cheaply (i.e., a supply cost of about $0.09/allele) and quickly using small-scale PCR reactions and UV plate readers.

Example 3 - Genetic tracking with IDP markers

Indel polymorphism (IDP) markers were successfully developed for the genetic tracking of particular alleles of several genes in segregating families. In some cases, the IDPs were as small as just a few basepairs (bps) in length.

The rates of IDP identification from a variety of types of DNA sequences were compared. Specifically, B73 and Mo17 sequences from 120 loci (or parts thereof) were analyzed. The rates of IDPs discovered between B73 and Mo17 in the three sources of DNA sequences were compared (Table 1).

Table 1. Frequencies of IDPs between B73 and Mo17 alleles.

	5' UTRs	Introns	3' UTRs
X(Y)/Z	8(2)/22	15(8)/54	10(3)/44

X= the number of sequences with at least one deletion in at least one allele

Y= the number of sequences with at least one deletion in both alleles

Z= the number of loci analyzed

About a third of the 5' UTRs and about a fourth of the 3' UTRs and introns examined from B73 or Mo17 have at least one IDP that can be used to design an IDP marker.

Example 4 - IDP Development

To exploit the high frequency of indels in maize introns, a large collection of

robust, allele-specific genetic markers for the high-throughput analysis of the maize genome is developed. A collection of about 1000 PCR primer pairs that reveal IDPs in corresponding alleles of the inbred lines B73, Mo17, the 16 inbred parents of BSSS, and the 12 inbred parents of BSCB1 are developed and genetically mapped. Because the maize genome is about 2000 cM, this number of IDPs provides, on average, one marker for each 2 cM. First, primer pairs (P1 and P2) from about 2000 pairs of exons are designed (Figure 4; panel A). These primer pairs are used to PCR amplify the introns that each exon pair flanks using genomic DNA from B73 and Mo17 as templates. Introns and primers are selected such that the resulting PCR products are about one kb in size. The resulting PCR-amplified intronic fragments are purified from agarose gels for each primer pair that yields a "clean" PCR product under a standard set of conditions. Both ends of each PCR product are sequenced using the two primers that were used during the amplification step (i.e., P1 and P2). Allele-specific primers (P3 and P4) are designed based on the IDPs identified between corresponding introns of B73 and Mo17 (Figure 4; panel B). Each pair of IDP primers consists of an allele-specific and a non-specific (exonic) primer, and is tested for specificity as illustrated (Figure 4; panel C).

The resulting IDP markers are genetically mapped using 350 RIs from the IBM population. The corresponding alleles from 1000 IDP loci that are well spaced across the genetic map are PCR amplified and sequenced from the 16 inbred parents of BSSS and the 12 inbred parents of BSCB1. Allele-specific primers are designed for each IDP locus. It is understood that more than one inbred may carry the same IDP allele at some loci and that it may not be possible to design allele-specific primers for all alleles. However, extremely useful IDP markers for most loci are generated. It is important to confirm empirically the allele specificities of each IDP primer pair under standard PCR conditions. For the purposes of identifying genes involved in heterosis, about 400,000 PCR reactions are conducted according to the following equation: [1000 (IDP loci) x 16 (primer pairs) x 16 (BSSS inbreds)] plus [1000 (IDP loci) x 12 (primer pairs) x 12 (BSCB1 inbreds)]. However, it is desirable to characterize fully the allele-specificities of the IDP primer sets. Thus, the allele-specificities of 30 allele primers from each of 1000 IDP loci are tested using a 30 (primer pairs) x 30 (inbreds) PCR array (i.e., 900,000 PCR reactions). To develop

IDPs, this strategy requires knowledge of gene structures and specifically the sequences of pairs of exons that flank introns. These data are available in GenBank for, at most, only a few hundred maize genes. Additional gene sequences are obtained from the maize genetics community as well as maize genic sequence databases generated via various plant genome projects (e.g., the Stanford-, Rutgers-, and Cold Spring Harbor-based maize genome projects). Once additional sequences are obtained, candidate introns are identified from predicted genes using Volker Brendel's maize-trained splice predictor program (SplicePredictor, available at http://gremlin1.zool.iastate.edu/cgi-bin/sp.cgi). IDPs also are generated from RFLP plasmids as described in Example 2. In addition, IDPs are identified using the sequences of 3' UTRs that are obtained according to Examples 1, 2, and 4, since 3' UTRs are also a rich source of IDPs.

Software to automate the computational steps required in IDP development is developed. This software (1) designs PCR primers to amplify intronic sequences, (2) assembles the "forward" and "reverse" sequences from the PCR-amplified introns, (3) confirms that the sequenced PCR-amplified intronic sequences are derived exclusively from the target gene sequence, (4) conducts multiple sequence alignments and identifies IDPs, and (5) designs PCR primers that are expected to be allele-specific based on these IDPs. Multiple alignments are conducted in a novel fashion. Heuristic algorithms for alignments based on a hydrophobicity index, residue coding, or other sequence variables are used to obtain initial alignments. Genetic algorithm-based alignment "polishing" software then is used to improve alignments. This technique should rival expert hand alignments for quality.

GenBank is an ideal source of maize DNA sequences from which to design primers for IDP discovery. Computer aided searches are used to identify records having desired information (e.g., maize DNA). GenBank records tend to be human readable but have formatting irregularities that require preprocessing before the records can be used in a high-throughput bioinformatics stream. To preprocess these records, software is designed and used to extract the necessary information and arrange the extracted information in a desired format. For example, software can be used to identify and extract introns from paired genomic DNA and cDNA sequences. To obtain the data shown in Table 1, PCR primers that amplify 5' UTRs, introns, and

3' UTRs were designed using software that analyzed and compared sequence and identified regions containing an IDP. About three-fourths (35/46) of these primers yielded single DNA fragments following PCR.

ESTs sequenced from their 3' ends are a readily available source of 3' UTR sequences. However, the amplification of 3' UTRs from such ESTs involves a primer design challenge. Ideally, one primer will be immediately 5' of the polyA site and the other some greater distance 5' of the polyA site such that the entire 3' UTR is amplified, but not so far 5' that coding region is included in the resulting PCR product. Since the sequence upstream of the stop codon of ESTs sequenced from their 3' ends can be of poor quality, it is not always possible to determine the 5' most stop codon and hence the beginning of the 3' UTR. To solve this problem, 76 different maize records from GenBank were analyzed to determine the size distribution of 3' UTRs. Results were grouped into bins of 50 base widths (Table 2). Based on these results, 5' primers were designed to hybridize about 400 bp 5' of the polyA site.

Table 2. 3' UTR length distribution in maize

Length of 3' UTR (bp)	Number of maize records	Length of 3' UTR (bp)	Number of maize records
0-50	1	450-500	0
50-100	0	500-550	1
100-150	3	550-600	0
150-200	17	600-650	1
200-250	18	650-700	0
250-300	13	700-750	0
300-350	12	750-800	0
350-400	5	800-850	0
400-450	4	850-900	1

Gene duplications are an important consideration in the design of PCR primers. They can complicate many experiments, including the PCR-based ("gene machine") system used to conduct reverse genetics in *Mu*-transposon containing

maize stocks. In an effort to alleviate these problems, primer design software is developed that interacts with available maize gene sequence databases to automatically design PCR primers with defined target specificities. In aid of this, the primer design tool incorporates efficient multiple sequence alignment tools. If primers are needed that will amplify only a specific gene or allele (as will be true with IDP design), then regions of the alignment that allow the target gene to be distinguished from its paralogs (or other known alleles) are used. In contrast, if it is desired to amplify every member of a particular gene family or all alleles of a gene, then regions of multiple alignments that exhibit sequence conservation are selected as target sites for primer design. This software includes an embedded knowledge base that gathers the results of previously conducted PCR experiments. By building a data acquisition routine into the primer design tool, it is possible to generate valuable training data. By mining this training set using adaptive algorithms, it is possible to induce new empirical rules to enhance primer design. As performance data accumulate, these rules will continue to improve. Such mining methods are designed to use genetic algorithm and genetic programming techniques such as those described elsewhere (Goldberg DE, Genetic Algorithms in Search Optimization, and Machine Learning. Addison-Wesley Publishing Company, Inc., Reading MA (1989); Koza J, Genetic Programming. MIT Press, Cambridge, MA (1992)).

Example 5 - Sequencing cDNA clones

To test the steps required to obtain nucleic acid sequence information for large numbers of clones, the following study was performed. A cDNA library was constructed from the inbred line B73. Templates were prepared using 96-well format Qiagen kits. Sequences were obtained from the 5' ends of 450 clones. In addition, the 3' ends of 62 of these clones were sequenced using a polyT(G/C/A) primer (PTN) that anneals to polyA tails.

Example 6 – Generating pooled libraries

Isolated mRNA from 20 different samples, including those from diverse seedling organs and developing kernels, was used for cDNA library construction. Other samples such as those from reproductive structures, those from maize seedlings

treated with gibberellic acid, cytokinin, ethylene, abscisic acid, auxin, bassinolide, and/or jasmonate, and those from maize calli treated with cycloheximide can be collected and used as well.

The cDNAs that are synthesized from different mRNA samples are combined into a single library. Unique tags are used to indicate the origin of individual cDNAs. These tags are added downstream of the polyA tail during the reverse transcription of each individual mRNA within a sample. A computer is used to generate large sets of sequence tags that are a specified number of insertions, deletions, and substitutions distant from one another such that mutations that occur during DNA replication do not confuse identification of the origin of a particular cDNA.

Example 7 - 3' EST Sequencing

To overcome the confounding effects of duplicate genes and retrotransposons, 3' UTR-enriched PCR products are generated for use in array-type experiments (e.g., MicroArray experiments). Although it is likely that some 3' UTRs contain retrotransposons, any sequences that contain recognizable retrotransposons are excluded from the array. A collection of 50,000 ESTs clones in microtiter dish format as bacterial cultures is obtained. These clones are picked into a 96-well format culture system using a Bio-robot. For long-term storage, clones are re-arrayed from the 96-well format into 384-well microtiter dishes that contain media, freezing solution, and the appropriate antibiotic. Sequencing templates are purified using 96-well format Qiagen kits. To determine the sizes of the inserts in these clones, the restriction digest products from each EST clone is subjected to low-resolution (i.e., high-throughput) electrophoresis. Sequencing is performed on an ABI3700 instrument. The sequencing of cDNA clones derived from polyT-primed libraries are performed using a polyT(G/C/A) primer (PTN) that anneals to polyA tails. Base calling is improved via the use of PHRED software. Remnant plasmid template DNAs not required for sequencing are placed into long-term storage to serve as templates for subsequent PCR reactions.

Rule-based adaptive computing methods and learning algorithms are used to flag suspicious DNA sequences. Sequences that are judged to have been incorrectly included or corrupted are saved in a library of errors for use by the adaptive error

checking routines. Examples of the types of checks made at this point include detection of vector sequence in the sequence interior (a type of chimeric sequence) or large frequencies of uncalled bases (N's). The system alerts the sequencing group if sequence quality falls below a specified minimum.

Since maize ESTs are a rich source of simple sequence repeats (SSRs), any SSRs found in the EST sequences are flagged as such. Based on SSR extraction experiments, it is predicted that as many as 25,000 candidate SSR sequences will be identified among the 50,000 3' EST sequences. Computer software can be used to locate both standard and imperfect SSRs based on the known properties of SSRs. The software required to accomplish these tasks is designed to handle streams of sequence data and can be revised to search incoming or all available sequences for any information the biological investigators deem interesting. All functions are performed automatically on batches of sequences as they are provided from the nucleic acid sequencing facility.

In addition, the 5' EST sequences are clustered into contigs. Because single-pass EST sequences usually contain base-calling errors, it will sometimes be difficult using only 5' sequence data to distinguish between cDNA clones derived from the same gene and those derived from closely related paralogous genes such as the *gl8a* and *gl8b* genes that are 97 percent identical in their coding regions. Among such paralogs, a higher level of DNA sequence polymorphism is usually observed in the 3' UTRs than in the coding regions. This natural source of useful information is exploited by conducting comparisons among the generated 3' EST sequences to help prevent the creation of artificial/chimeric EST contigs. For these DNA sequence analyses, the efficiency of standard techniques (e.g., using BLAST to match query sequences to sequences in a database) is compared with a new technique. The new technique fragments gene sequences into a dictionary of short subsequences that contain not only those short subsequences but also the number of times each subsequence are encountered. For any two such dictionaries, a homology number is generated by treating the dictionaries as vectors and computing the angle between them. In addition to being roughly as fast as BLAST for pair-wise sequence comparisons, this new technique can, by merging dictionaries, compare an EST to a cluster of sequences in a single pass. This latter capability permits a speed increase

when placing ESTs into a clustered database. This increase is speed is roughly proportional to the average cluster size in the database.

Once EST clusters have been generated, a comparative genomics study is conducted using nucleotide and predicted amino acid sequences. A hierarchy of gene families is built using phylogenetic analysis to distinguish the major gene clusters. This gene hierarchy is subsequently refined to define the interrelationships among related genes arising from recent gene duplications. Three data sets are analyzed: the 3'-EST clusters, 5'-EST contigs from a maize genome project, and the combined EST data set. The predicted amino acid sequences of the combined EST data set are suitable for identifying ancient gene families, whereas 3'-ESTs (which include 3' UTRs) may have higher statistical resolution for resolving recent duplications since 3' UTRs contain more sequence variation than coding regions. These clustering efforts will define a set of non-redundant EST clones that are used to generate targets for array experiments.

For each EST cluster that was generated by one or more recent duplications, potential recombination and/or gene conversion events are detected by comparing the separate phylogenetic trees inferred from the sequences of 3' or 5' ESTs. Once the genetic map positions of these EST clones are established, a genome-wide picture of the patterns of gene duplications that have occurred during maize evolution is developed. For example, the fate of a large gene family clustered in a local region can be studied. Since this type of gene family can be related to important physiological processes (e.g., disease-resistance), the interaction between genome doubling and adaptation to environment is addressed. The DNA sequence and predicted proteins of each EST contig is compared to the non-redundant GenBank database and cross-linked to GenBank, MaizeDB (http://www.agron.missouri.edu/), and ZmDB (http://www.zmdb.iastate.edu/). In addition to serving as a rich source of SSRs, this collection of 3' EST sequences, in combination with genomic DNA sequences being generated by other maize genome projects, provides data useful to others in defining maize polyadenylation signals.

Example 8 - PCR Amplification of 3' UTRs

A set of about 10,000 non-redundant ESTs is selected using the data generated

in Example 7. The 3' sequence of each of these EST clones is PCR amplified using gene-specific (GS) and PTN primers. The primer design tools described herein are used to automate the primer design steps. Because the gene-specific primers are designed based on sequences about 300-400 bp 5' of the polyA tails in the 3' EST sequences, the resulting PCR products are enriched for 3' UTRs. The resulting PCR products are used as targets for array-type experiments described herein. Figure 5 depicts the 3' fragments that were PCR amplified in this manner from 29 random ESTs clones for which 3' sequences were obtained.

Example 9 – Specificity of 3' UTR versus full-length EST sequences

To determine whether 3' UTRs provide a greater degree of gene specificity than full-length EST clones, the following experiment was performed. First, PCR primers were designed and used to amplify the 3' UTRs of a collection of 192 EST clones. Second, the cDNA inserts of the 192 clones were PCR amplified. The resulting 384 PCR products were arranged on an array based on that of replicated field plot experiments. The degree of gene specificity is typically determined by hybridization of the array with, e.g., mRNA from a particular cell. Essentially equivalent hybridization to both 3' UTRs and full-length EST clones is an indication of very little gene specificity, while differential hybridization to, for example, 3' UTRs, indicates a greater degree of gene specificity for 3' UTR sequences.

Example 10 - Mapping Array Probes and Targets

In a traditional Southern blot, a radioactively labeled *rf2a* cDNA probe detected a *Hin*dIII RFLP between the inbred lines Co159 and Tx303 (Figure 6). To map the *rf2a* gene using this current technology, a hybridization of this type would be conducted using DNAs from a mapping population segregating for this RFLP. To map a second gene using this approach, a second probe would need to be synthesized and another hybridization conducted. Thus, it would be necessary to conduct 10,000 labeling reactions and hybridizations to map 10,000 genes using this technology. Even more seriously, this technology requires a large number of time-consuming and expensive electrophoresis procedures and Southern blot transfers.

In the Southern blot experiment, the genomic DNA served as the "target" and

the cDNA as the probe. As described herein, a mapping technology that overcomes the throughput limitations inherent in current RFLP-based mapping approaches was developed. This technology generates a genetic map containing about 10,000 cDNAs resulting in a genetic map with an average density of five genes per cM. In this new procedure, the cDNA clone serves as the target (on an array) and the probe consists of size-fractionated genomic DNA from the mapping population.

A primary challenge that was overcome involved obtaining probes in which the genic sequences from the size-fractionated genomic DNAs had sufficient specific activities to hybridize to the cDNA targets. First, genomic DNA was digested with *Hin*dIII and size-fractionated via electrophoresis through an agarose gel. This gel was then sliced into serial fractions each of which contained about 5 percent of the total maize genome. Aliquots of purified size fractions of genomic DNA from the inbred line Co159 were subjected to electrophoresis (Figure 7).

The remaining aliquots of each size fraction were sonicated briefly, denatured by boiling, and then allowed to reanneal at 68°C to a CoT value of 4.8 (Zwick *et al.*, *Genome* 40:138-142 (1997)) prior to being radioactively labeled. This reannealing step removed much of the repetitive DNA from the labeling reaction, thereby increasing the labeling of the genic sequences from the gel slices. As demonstrated by the traditional Southern blot analysis (Figure 6), the 4.5 to 5 kb size fraction from Co159 contained the *rf2a* gene. When this fraction was labeled as described, it hybridized to a nylon membrane Southern blotted with a fragment of the *rf2a* cDNA (Figure 8; lane 1), but not to a fragment from another maize cDNA (Figure 8; lane 2). The 5.5 to 6 kb size fraction from Co159 did not contain the *rf2a* gene. When similarly labeled it did not hybridize to a near-identical membrane containing a fragment of the *rf2a* cDNA.

The specificity of these size-fractionated genomic DNA probes was further demonstrated by the fact that the 9 to 10 kb size fraction from Co159 hybridized to the PCR-amplified 3' region of only one of 29 random EST clones (Figure 5; lane 27 on right panel). Thus, these results demonstrate the feasibility of using this approach to map genes. If two RI lines from a mapping population carry different RFLP alleles of a given gene (e.g., 4 kb and 6 kb, respectively), then the 4 kb, but not the 6 kb fraction of RI #1 will hybridize to the corresponding cDNA target on the array. In contrast,

the 6 kb fraction, but not the 4 kb fraction of RI #2 will hybridize to this target. Thus, if a series of arrays containing 10,000 non-redundant sequenced gene clones is hybridized with fluorescently labeled genomic DNA size fractions from each individual in a mapping population, it will be possible to map simultaneously many of the 10,000 genes.

Example 11 - Mapping Arrays

An Arrayer instrument is used in conjunction with the 3' ends of about 10,000 non-redundant, sequenced gene clones to produce arrays (i.e., a mapping array). This collection of clones is selected such that it contains some of the several hundred maize cDNAs that have previously been genetically mapped in maize. These controls serve to anchor the resulting map relative to existing maize genetic maps.

A mapping array is hybridized with fluorescently labeled, serial, size-fractioned genomic DNA from individual maize RI lines from the IBM mapping population. Fluorescent signals are detected with a General Scanning ScanArray instrument.

A significant experimental design question is how to maximize the efficiency of a mapping experiment with the minimum number of probes and chips. The number of chips is dependent upon the number of probes that can be simultaneously hybridized to a given single-use chip. The current General Scanning instrument detects only two fluorescent signals (Cy3 and Cy5); however, the next generation of General Scanning instruments (ScanArray 5000) is capable of detecting four fluorescent dyes per chip. Using the ScanArray 5000 instrument, each hybridization includes labeled DNA fractions from two RI lines and both parental controls. The presence of both parental hybridization signals on each chip serves as a control in genotyping the two RI lines. Using the current General Scanning instrument, a two-channel detector is used, and the same quality of data is collected from four chips, each hybridized with labeled DNA fractions from one RI and one parent.

Since the number of probes needed is the product of: (# restriction enzymes used to digest the probe DNA) x (# gel-slices per RI line) x (# RI lines included in the mapping panel), the experimental design must consider each of these variables.

1. Number of restriction enzymes

Not every marker will be polymorphic following digestion with a single enzyme. If DNA is digested with a second enzyme, some proportion of markers that were monomorphic when analyzed with the first enzyme will now be polymorphic and thus mappable. Let p denote the proportion of markers that exhibit a polymorphism for any given enzyme. Assume that p is constant across enzymes, and that the polymorpic-monomorphic status of a marker using a given enzyme is independent of its status for all other enzymes. It follows that the proportion of markers that are polymorphic for at least one enzyme should be well approximated by $1-e^{\lambda x}$, where x is number of enzymes and $\lambda = -\log(1-p)$. These expectations were tested using an empirical polymorphism survey from the IBM population containing over 100 RFLP markers and five restriction enzymes. Based on these data, it was predicted that 60 to 70 percent of markers will be mappable when only a single restriction enzyme digest is employed. However, these data also were used to demonstrate that there is a positive correlation among markers across enzymes. Thus, if a marker is monomorphic for enzyme A, then it is slightly more likely than random to also be monomorphic for enzyme B. Therefore, a second enzyme increases the proportion of polymorphic markers to about 80 percent, while four enzymes may increase the proportion of polymorphic markers to about 90 percent. Consequently, at least two but not more than four enzyme digests are used in a mapping array.

2. Number of gel slices

Gel slices define bins; polymorphisms can be detected only if two parental bands land in distinct bins. Thus, markers that would have been mappable using standard Southern blots will appear monomorphic if they are included in the same bin. To determine how much impact this will have on the proportion of mappable markers is a function of the size of the band shift associated with polymorphic markers. This question was examined empirically using existing data, but it seemed reasonable to assume that most band shifts are approximately exponential in their length distributions. If the gel positions of the bands from parents A and B are denoted as x_A and x_B, then the band shift $\delta = x_A B x_B$ is distributed as a double-exponential with two of the densities being positive and two negative. If the gel were cut into uniform slices

of size ω, then calculations demonstrate that the proportion π of polymorphic markers that are detectable by the sliced-gel approach can be expressed as a function of the slice width ω and the mean shift size μ. Thus,

$$\pi = 1 - \frac{\mu}{\omega}\left(1 - e^{-w/\mu}\right)$$

Using this relationship, it was observed that: (1) if ω = μ, 50 percent of the polymorphic markers can be detected; (2) if ω = μ/2, 80 percent of polymorphic markers can be detected; and (3) if ω = μ/5, 90 percent of polymorphic markers can be detected. The diminishing returns of slicing more and more finely are clear and the optimal number of gel slices can be determined.

3. Size and composition of the mapping panel

The mapping panel consists of a subset of size *n* RI lines, selected from the available 350 in the IBM population. Because the goal is to ensure that the mapping panel contains a sufficient number of recombination breakpoints to map the 10,000 ESTs, two variables need to be considered: the size of the mapping panel and its composition (i.e., which RIs are included in the panel). By careful selection, the information content of a panel of size *n* is maximized such that the resolution obtained will be greater than a panel composed of *n* random lines.

A maximally informative mapping panel would have a large number of regularly spaced recombination breakpoints along the chromosomes. Lines are selected from the IBM mapping population to be included in the mapping panel based on their genotypes as revealed with a subset of all markers (i.e., screening markers). Genotypic data on these RIs for >400 RFLP and SSRs markers were obtained. In addition, similar genotypic data is obtained for about 1000 IDP markers. Only those markers for which high-quality data (i.e., with low rates of missing data and double crossovers/potential errors) are available are used for this selection.

To identify a highly informative subset of lines for the mapping panel, an optimality criteria *U* is calculated for each candidate subset. Because the number of possible subsets is huge, a Monte Carlo optimization (e.g. simulated annealing) and/or greedy approach (serial additions of the next best line) is used to obtain good

candidate panels. The criterion U is computed for these panels until a reasonably optimal panel is identified.

The optimality criterion is the mutual entropy of marker genotypes. Because chromosomes segregate independently, the entropies can be computed chromosome-wise and summed. Thus,

$$U = \sum_{h=1}^{20} U_h$$

where U_h is the entropy for chromosome h. Now suppress the chromosome indexing and consider a single chromosome with m screening markers. If a no-interference model is assumed for recombination, the genotype of a chromosome (a series of marker typings, e.g., AAAABBBBBA) forms a Markov chain, call it $\{X = x_1,...x_n\}$. The entropy of a Markov chain is

$$U = E(-\log(P(X)))$$

$$= -\sum_{s\in\{A,B\}} P(X_1=s)\, P(X_1 = s)$$

$$-\sum_{i=1}^{m} \sum_{s\in\{A,B\}} \sum_{t\in\{A,B\}} P(X_{i-1}+s, X_i + t)\ \log P(X_i = t | X_{i-1} = s).$$

The probabilities (initial state, joint and conditional) are simply estimated from data, i.e., from marker genotypes on the candidate panel. Estimates of probabilities are then substituted into the expression to obtain an estimate of the entropy.

The optimum value of n (the panel size) is identified by conducting simulation experiments in which the resolving powers of panels of different sizes are determined. Suppose a collection of M=10,000 markers needs to be mapped and that the panel has been constructed with a screen set of m=100 markers. Assuming that these markers are distributed evenly across the genetic map, a typical interval between screening markers will contain M/m = 100 markers to be mapped. A critical quantity that will be determined empirically using the screening marker data set is the mean number of breakpoints (summed across the mapping panel) per interval in the screening map. The fact that the IBM population was intermated for several generations prior to the extraction of inbred lines helps increase this quantity. Once these quantities are

available, the results of Thompson (*IMA J. Math. Applied Med. Biol.*, 1:31-49 (1984)) are applied to help estimate the resolving power of the mapping panel. With r randomly placed recombination events and m markers to be mapped, the expected proportion of "isolated" markers is $q= r(r+1)/(r+m)(r+m+1)$, and the expected number of bins = $(r+1)(1-q)$. Thus, by mapping 100 markers (m=100) with 100 breakpoints (r=100), it would be expected that 1/4 of the markers will be isolated (i.e., will have breakpoints on either side of it that separates it from all other markers) and the set markers will be divided into 25 bins. With 200 breakpoints (r=200), 4/9 of the 100 markers can be isolated into 45 bins. Thompson (*IMA J. Math. Applied Med. Biol.*, 1:31-49 (1984)) provides variances for these quantities that will be useful in making a more detailed analysis of the resolving power. Current calculations suggest that 100-200 RIs is a reasonable target.

Given that it will be possible to determine which gel slices contain the B73 and Mo17 alleles (because B73 and Mo17 will be included as controls in the mapping panel) and Mendelian segregation can be assumed, a fully Bayesian classifier can be built based on a linear model that could account for important sources of variation in hybridization signals. This would be ideal and would provide posterior probability of presence/absence for each RI line x gel-slice x enzyme combination.

Because the parental origin of each mappable marker will be known, it is possible to use standard mapping approaches. However, some potential complications may require the development of custom mapping software. A multiple imputation method for error correction and missing genotypes was developed for this project. In this context, errors and missing data are handled simultaneously. The procedure is flexible in that it can allow for error rate heterogeneity across markers and asymmetry in the error process. The approach is to impute several versions of complete and corrected data sets, and to analyze the ensemble of that data to produce a final map. The procedure is computationally efficient and provides measures of uncertainty that are not readily available otherwise.

If multiple enzymes are used for genotyping, a number of markers will be polymorphic for more than one enzyme and redundant genotypes will be obtained. Thus, genotypes may be determined with higher accuracy for some markers than for others. If the duplicate reads are concordant, there is no problem. However,

discordant genotypes will almost certainly be obtained in some instances. These data are used to estimate the rate of genotyping errors. Assuming that double errors are very rare (i.e., markers are not mistyped using two enzymes), the numbers of discordant and concordant genotypings is determined among those that are redundant. Existing data can be used to determine the optimal values for each of the three parameters. Several hundred probes were hybridized to DNA gel blots containing B73 and Mo17 DNA digested with a variety of restriction enzymes. The resulting RFLP patterns can be used to determine the optimal number of enzymes and number of gel slices per inbred.

About 50 RFLP probes were analyzed. For each probe, the size of hybridizing DNA fragments following digestion with each of the restriction enzymes was recorded for each inbred. In addition, 350 IBM RIs were genotyped with a large number of RFLP probes. These data, and those generated with IDPs, can be used to identify particular RIs.

Example 12 - Mapping of mutants

Maize biologists have accumulated a vast collection of single-gene mutants that confer a diverse spectrum of phenotypes that affect traits of biological and agricultural interest (Neuffer *et al.*, Mutants of Maize. Cold Spring Harbor Laboratory Press, Plainview, New York (1997)). The analysis of these mutants is greatly facilitated by genetically mapping the affected genes relative to molecular markers. For example, the availability of linked genetic markers simplifies the generation of specific genotypes (needed, for example, to create double mutants and conduct enhancer or suppressor screens) and allows candidate gene cloning experiments to be conducted. Although unique cytogenetic stocks are available for mapping maize mutants defined only by phenotype (e.g., BA and waxy-marked AA translocation stocks), such mapping experiments are laborious and time-consuming. In addition, mapping with cytogenetic stocks is difficult for traits that exhibit epistatic interactions.

To conduct a mapping experiment of this type, it is necessary to have available a population that is segregating for the mutant of interest. Various population structures can be used, but to illustrate the procedure, backcross and F_2 populations

segregating for the recessive mutant *a* and the wild-type allele *A* are used. Backcross mapping populations are derived from the cross: *a/A* X *a/a* and will segregate 1:1 for mutant (*a/a*) and wild-type (*a/A*) individuals. F_2 populations are derived from the self-pollination of heterozygous (*a/A*) individuals, and will segregate 1:3 for mutant (*a/a*) and wild-type (*a/A* and *A/A*). The mapping array is used to identify those polymorphic loci that exhibit a bias in allele distribution between mutant and wild-type plants. This is accomplished by creating pools of DNA from the two phenotypic classes (i.e., bulked segregant analysis; Michelmore *et al.*, *Proc. Natl. Acad. Sci. USA*, 88:9828-9832 (1991)). These two pools are digested with several restriction enzymes and subjected to gel electrophoresis. Paired sets of size fractions are purified from the two DNA pools. Paired size fractions from the two pools are labeled with Cy3 and Cy5, respectively, and hybridized to an array.

The Cy3 and Cy5 signals (representing the mutant and wild-type DNA pools) are equal for cDNAs on an array that are derived from loci that are not closely linked genetically to gene *a*. In contrast, the Cy3 and Cy5 signals of loci that are closely linked to gene *a* will exhibit signal biases. The intensity of the bias at a given marker locus will be inversely proportional to its linkage to gene *a*.

To test the feasibility of using a mapping array to map genes defined only by phenotype, a subset of a mutant collection is analyzed. A total of ten genes are initially mapped. Mapping populations are available for five newly defined "glossy" genes, four "root hair" genes, and *pif1*. Glossy and root hair genes are those that when mutated, alter the accumulation of cuticular waxes on seedling leaf surfaces or the development of root hairs, respectively. The *pif1* gene interacts with an aldehyde dehydrogenase (encoded by the *rf2* gene) to affect male fertility. To demonstrate the validity of the resulting mapping data obtained with a mapping array and to calibrate the relationship between Cy3:Cy5 signal bias and genetic distance, the genetic map positions obtained for these ten genes are confirmed using standard RFLP analyses.

Example 13 - Identification of Genes involved in Heterosis

For the purposes of this invention, genes that affect heterosis (heterosis genes) are divided into two types: *cis* heterosis genes (CHGs) and *trans* heterosis genes (THGs). Favorable alleles of CHGs are directly responsible for elevated levels of

heterosis. As such, the frequencies of favorable CHG alleles are expected, on average, to increase during the course of selection for heterosis and yield. THGs are those genes that exhibit altered levels of gene expression in hybrids (relative to the corresponding inbred parents) and in some instances may be regulated by CHGs. The allele frequencies of THGs may or may not have changed during the course of the RSS program. Note that this classification system (CHG versus THG) makes no assumptions as to the regulatory versus structural natures of these genes.

Two related approaches are used to identify a subset of the CHGs and THGs that play a role in heterosis in the BSSS and BSCB1 populations. As the first step in identifying CHGs, the IDP markers generated as described herein are used to identify those chromosomal intervals that have experienced the largest changes in allele frequencies during selection for yield and heterosis in the BSSS and BSCB1 populations. Based on the EST mapping data obtained as described herein, it is possible to identify candidate CHGs within these chromosomal intervals. The effects of these chromosome intervals (and the candidate CHGs) on heterosis and gene expression are assayed in replicated yield trials and using array technology, respectively. This latter test will identify THGs.

IDPs are used to identify those chromosomal intervals whose allele frequencies have increased in response to selection for heterosis and yield in the BSSS and BSCB1 populations. This is accomplished by genotyping the 16 inbred progenitors of BSSS and the 12 inbred progenitors of BSCB1 to represent the Cycle 0 population (base population), 75 plants from the Cycle 5 and 9 populations, and the 20 progenitors of Cycles 11 and 14 with 250 of the most informative of the 1000 IDP markers developed as described herein. Thus, a total of 206 plants are genotyped from BSSS and 202 plants from BSCB1. Only the nature of IDP markers makes an analysis of this magnitude [(206 x 250 x $\leq$16) + (202 x 250 x $\leq$12) $\leq$ 1,430,000 PCR reactions] possible. The small-scale PCR reactions are conducted in 96-well microtiter plates and data directly collected with a plate reader, or, alternatively, very high-throughput, capillary-based PCR "chips" are used (Koop *et al.*, *Science*, 280:1046-1048 (1998)).

After genotypic data are collected, population genetic parameters are analyzed using public software available at http://evolution.genetics.washington.edu/. Software

such as Genepop (http://www.ualberta.ca/~fyeh/index.htm) and/or GDA (http://alleyn.eeb.uconn.edu/gda) are used to summarize allele frequency data and provide estimates of population genetic statistics. Waple's statistical tests of directional selection based on temporal changes in allele frequency also are applied to these data.

The ten chromosomal intervals that exhibit the greatest changes in allele frequency are identified for each population. Subsequently, 200 existing random S4 inbred lines derived from Cycle 14 from each population are genotyped with 50 of the most informative IDP markers in the vicinity of these ten chromosomal intervals [(200 x 50 x ≤ 16) + (200 x 50 x ≤ 12) ≤ 280,000 PCR reactions)]. To test whether CHGs in these 20 chromosomal intervals actually affect yield or heterosis, each of the 200 S4 lines from BSSS are crossed by bulked pollen from BSCB1 (i.e., subjected to a topcross), and each of the 200 S4 lines from BSCB1 are similarly crossed by bulked pollen from BSSS. The 400 resulting topcross populations are yield tested in a replicated plot design (2 reps x 5 locations x 2 years).

For each of the ten candidate chromosomal intervals from BSSS and those from BSCB1, the topcross yields and percent mid-parent heterosis are compared for all S4 lines that carry the "favorable" allele of each candidate chromosomal interval with topcross yields of the remaining S4 topcrosses. In this context, "favorable" is defined as that allele whose frequency increased most significantly during 14 cycles of the RSS selection experiment. Those chromosomal intervals that confer statistically significant yield and percent heterosis differences on the topcross progeny that carry them are predicted to contain CHGs.

The arrays described herein are used to identify THGs and those CHGs that differentially regulate gene expression levels in hybrids. Samples of mRNAs from inbred parents and their respective hybrid progeny are converted to cDNA and labeled using fluorescent dyes. Detection is performed using a General Scanning ScanArray 3000 instrument that is capable of detecting two distinct fluorescent signals per slide. Alternatively, a General Scanning ScanArray 5000 instrument that can detect four distinct fluorescent signals per slide is used. In this case, the three-way comparisons (Parent 1, Parent 2, and hybrid) on a single slide are performed; otherwise, two slides are used for each three-way comparison. Initially, a large number of diverse tissues

and developmental stages from B73, Mo17, and the B73 x Mo17 F_1 hybrid is analyzed to identify the five tissues and developmental stages in which the maximum number of genes exhibit the largest differences in expression between the F_1 and parents.

The two chromosomal intervals from each population that are the best predictors of topcross progeny yield and percent heterosis will be identified from these experiments. Ten S4 lines from each population that carry and do not carry each of these chromosomal intervals are identified and their effects on gene expression determined. An array of 10,000 ESTs is used to assay gene expression at the five tissues/developmental stages identified above in the 40 S4 inbred lines (2 populations x 20 S4 lines), their 40 topcross progeny, and pooled plants from the two populations. Genes that exhibit alterations in gene expression in hybrids relative to the parents are classified as candidate THGs or CHGs. It is expected that both CHG-regulated and CHG-"independent" THGs will be identified. CHG-"independent" means not regulated by alleles of the two CHG-related chromosomal intervals under analysis from the two populations.

Using data from a cDNA mapping array experiment described herein, the genetic map positions of the identified CHGs or THGs are determined. The next step involves setting up a cycle of selection in which the effectiveness of selection based on alleles of candidate CHGs is compared to that based on progeny tests.

Pattern recognition algorithms are used to extract biological meaning from these data sets. The scientific community is still struggling with the best tools with which to extract such information. However, clustering genes that have related expression patterns in an effort to develop hypotheses regarding gene function can be used (Somogyi and Sniegoski, *Complexity*, 1:45-63 (1996); Carr *et al.*, *Statist. Comp. Statist. Graph. Newsletter* pp 20-29 (1997); and Wen *et al.*, *Neurobiol.* 95:334-339 (1998)). This is a reasonable approach because many processes in biology are Markovian and hierarchical. For example, all signal transduction cascades are hierarchical in that interactions late in a cascade cannot occur unless product derived from an earlier stage is present and available. Cladistic analysis affords a powerful means of visualizing latent hierarchical signals that exist in data sets. Traditionally, cladistic analysis has been employed to estimate the hierarchical relationships among

species in the form of evolutionary trees. However, the methodology can readily be co-opted to deduce the hierarchical temporal relationships among interacting gene products in a genome. Thus, gene products that appear hierarchically closely related are likely to belong to the same pathway.

OTHER EMBODIMENTS

It is to be understood that while the invention has been described in conjunction with the detailed description thereof, the foregoing description is intended to illustrate and not limit the scope of the invention, which is defined by the scope of the appended claims. Other aspects, advantages, and modifications are within the scope of the following claims.